Suresh Gyan Vihar University Journal of Engineering & Technology
(An International Bi-Annual Journal)
Vol . 3, Issue 2, 2017, pp 48-52
ISSN: 2395-0196

Analysis of Various Load Balancing Techniques in Cloud Computing: A Review

Jyoti Rathore
Research Scholar Computer Science & Engineering, Suresh Gyan Vihar University, Jaipur
Email: Jyoti.rathore131@gmail.com
Dr. Bright Keswani
Professor, Department of Computer Applications, Suresh Gyan Vihar University, Jaipur
Email: Kbright@rediffmail.com
Dr. Vijay Singh Rathore
Professor, Department of Computer Science & Engineering, Jaipur Engineering College & Research Centre, Jaipur
Email: Vijaydiamond@gmail.com

Abstract: **Cloud computing is a new trend emerging in IT environment with huge requirements of infrastructure and resources. Load Balancing is an important aspect of cloud computing environment. Efficient load balancing scheme ensures efficient resource utilization by provisioning of resources to cloud user's on-demand basis in pay-as-you-say-manner. Load Balancing may even support prioritizing users by applying appropriate scheduling criteria. This paper presents various load balancing schemes in different cloud environment.**
Keywords: **Cloud Computing, Load Balancing, Distributed computing, etc.**

I. INTRODUCTION

A Cloud computing is emerging as a new technology of large scale distributed computing [10]. Commercial cloud computing providers, such as Google, Amazon, Yahoo and Microsoft deliver cloud computing service to customers all over the world. Cloud computing is platform independent as there is no need to install software's in PC's it provide online applications to users on-demand. As cloud computing is in its evolving stage, so there are many problems prevalent in cloud computing. The most prevalent problem in Cloud computing is load balancing [2]. Load balancing is the mechanism of distributing the load among various nodes to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are idle or doing little work while other are heavily loaded [10]

II. CLOUD COMPUTING: DEFINITION

The " Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, data storage, software applications and other computing services) that can be rapidly provisioned and released with minimal management effort or service provider interaction "[4]. By sharing of resource the overall cost reduces. Cloud computing delivers all services through the internet dynamically when user demands, such as operating system, network, storage, software, hardware and resources [11]. Cloud is a pool of heterogeneous resources. It is a mesh of huge infrastructure and has no relevance with its name "Cloud". Infrastructure refers to both the applications delivered to end users as services over the Internet and the hardware and system software in datacenters that is responsible for providing those services. In order to make efficient use of these resources and ensure their availability to the end users "Computing" is done.

III. LOAD BALANCING: CLOUD COMPUTING NEED

An ideal load balancing algorithm should avoid overloading or under loading of any specific node [4]. But, in case of a cloud computing environment the selection of load balancing algorithm is not easy because it involves additional constraints like security, reliability, throughput etc. So, the main goal of a load balancing algorithm in a cloud computing environment is to improve the response time of job by distributing the total load of system. The algorithm must also ensure that it is not overloading any specific node. The important things to consider while developing such algorithm are the estimation of load, the comparison of load, the stability of different system, the performance of system, the interaction

between the nodes, the nature of work to be transferred, the selection of nodes and many other ones [4]. This load considered can be in terms of CPU load, amount of memory used, delay or Network load.

There are mainly two types of load balancing algorithms

1) Static Algorithm in which traffic is divided evenly among the servers. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static algorithm is proper in the system which has low variation in load.

2) Dynamic Algorithm in which the lightest server in the whole network or system is searched and preferred for balancing a load. For this real time communication with network is needed which can increase the traffic in the system. Here current state of the system is used to make decisions to manage the load.

IV. TECHNIQUES SUPPORTING LOAD BALANCING - A LITERATURE

The Round Robin is one of the simplest scheduling techniques that utilize the principle of time slices. Here the time is divided into multiple slices and each node is given a particular time slice or time interval i.e. it utilizes the principle of time scheduling. Each node is given a quantum and its operation. The Resources of the service provider are provided to the requesting client on the basis of time slice [17].

The max-min algorithm is much the same as to min-min algorithm. At first for all the available tasks are submitted to the system and minimum completion time for all of them are calculated, then among these tasks the one which is having the completion time, maximum is chosen and that is allocated to the corresponding machine. This algorithm outperform than Min-Min algorithm where when short tasks are in high numbers when compared to that of long ones. For e.g. if in a task set only a single long task is presented then ,Max Min algorithm runs short tasks concurrently along with long task. The make span focus on how much small tasks will get executed concurrently with the large ones. Max-Min is almost identical to Min-Min, except it selects the task having the maximum completion time and allocates to the corresponding machine. The algorithm suffers from starvation where the tasks having the maximum completion time will get executed first while leaving behind the tasks having the minimum completion time [8].

The Min-Min Algorithm take up with a task set which are initially not assigned to any of the nodes. Initially the minimum completion time is calculated for all the available nodes. Once this calculation gets completed the task having the completion time minimum is chosen and assigned to the respective node. The execution time of all other tasks which are currently available in that machine is updated and the task gets discarded from the available task set. The routine is done time after time until all the tasks have been assigned to the equivalent machines. The algorithm works better when the situation is like where the small tasks are greater in number of than the large tasks. The algorithm has a disadvantage that it leads to starvation. smaller tasks will get executed first, while the larger tasks keeps on in the waiting stage ,which will finally results in poor machine use [8].

The Ant Colony Optimization Technique is a pheromone table was being designed which was updated by ants as per the resource utilization and node selection. Ants move in forward direction in search of the overloaded or under loaded node. As the overloaded node is traversed, then ants move back to fill the recently encountered under loaded node, so a single table is updated every time [16].Ant algorithms is a multiagent approach to difficult combinatorial optimization problems. Example of this approach is travelling salesman problem (TSP) and the quadratic assignment problem (QAP). These algorithms were inspired by the observation of real ant colonies. Ant’s behaviour is directed more to the survival of the colonies .They not think for individual [24].

The Honey Bee algorithm is a nature inspired load balancing technique which helps to achieve load balancing across heterogeneous virtual machine of cloud computing environment and maximize the throughput. First the current workload of the VM is calculated, then it decides the VM states whether it is over loaded, under loaded or balanced. The priority of the task is taken into consideration after removed from the overloaded VM which are waiting for the VM. Then the task is scheduled to the lightly loaded VM. It reduces the response time of VM and also reduces the waiting time of task [13].

The Equally Spread Current Execution technique load balancer makes effort to preserve equal load to all the virtual machines connected with the data centre. This load balancer maintains an index table of Virtual machines as well as number of requests currently assigned to the Virtual Machine (VM). If

the request comes from the data centre to allocate the new VM, it scans the index table for least loaded VM. In case there are more than one VM is found than first identified VM is selected for handling the request of the client/node, the load balancer also returns the VM id to the data centre controller. The data centre communicates the request to the VM identified by that id. The data centre revises the index table by increasing the allocation count of identified VM. When VM completes the assigned task, a request is communicated to data centre which is further notified by the load balancer. The load balancer again revises the index table by decreasing the allocation count for identified VM by one but there is an additional computation overhead to scan the queue again and again [3].

The Active Clustering works on the concept where same type nodes of the system are grouped together and working on these groups. It works like as self-aggregation load balancing technique where network is rewired to balance the load of the system. Systems optimize using similar job assignments by connecting similar services. The performance of the system is enhanced with high resources thereby increasing the throughput by using these resources effectively [2]. As number of nodes increase, Active Clustering and Random Sampling Walk predictably performed better although the latter resulted into dramatic variation in overall performance [6].

The Biased Random Sampling is based on the construction of the virtual graph having connectivity between the all nodes of the system where each node of the graph is corresponding to the node computer of the cloud system. Edges b/w nodes are two types as Incoming edge and outgoing edge that is used to consider the load of particular system and also allotment the resources of the node. It is scalable technique to balance the load of the cloud system. It is also reliable and effective load balancing approach that is mainly developed to balance the load of distributed system. Load balancing is achieved without the need to monitor the nodes for their resources availability [2]. As number of nodes increase, Random Sampling Walk predictably performed better although the latter resulted into dramatic variation in overall performance [6].

In throttled algorithm the load balancer maintains the record of each state (busy or ideal) in an index table of virtual machines. First the client or server makes a request to data center to find a suitable virtual machine to perform the recommended job. The data center queries the load balancer for allocation of the VM. The load balancer check the index table from top until the first available VM is found, if the VM is found, the load data center communicates the request to the VM identified by the id. Further, if appropriate VM is not found, the load balancer returns -1 to the data center. When the VM completed the allocated task, a request is acknowledged to data center, which is apprised to load balancer to de- allocate the same VM whose id is already communicated. The total execution time is estimated in three phases. In the first phase the formation of the virtual machines and scheduler will be idle waiting to schedule the jobs in the queue ,in second phase once jobs are allocated, the virtual machines in the cloud will start processing, and finally in the third phase the destruction of the virtual machines. The throughput of the model can be estimated as the total number of jobs executed within a required time span without considering the any destruction time. This algorithm will improve the performance by providing the resources on-demand, by reducing the rejection in the number of jobs submitted and resulting in increased number of job executions [10].

The Map Reduced based Entity Resolution load balancing technique is based on large datasets. There are two main tasks: Map task and Reduce task. The PART method is executed for mapping task, where the request entity is partitioned into parts. And then COMP method is used to compare the parts and finally GROUP method is use to group similar entities and then similar task are group using Reduce function and this also reduces the results of the tasks. Map task reads the entities in parallel and process them, so that overloading of the task is reduced [2].

The Opportunistic Load Balancing (OLB) is a static load balancing algorithm whose goal is to keep each node in the cloud busy so does not consider the current load on each node. It attempts to dispatch the selected job to a randomly selected available VM. However, OLB does not consider the execution time of the task in that node. This may cause the task to be processed in a slower manner increasing the whole completion time (makespan) and will cause some bottlenecks since requests might be pending waiting for nodes to be free [15].

The Decentralized Content Aware is a new content aware load balancing policy named as workload and client aware policy (WCAP) that uses a unique and special property (USP) to specify the unique and

special property of request as well as computing nodes. USP helps the scheduler to decide the best suitable node for processing the requests. This strategy is implemented in a decentralized manner with low overhead. By using the content information to narrow down the search, it improves the searching performance overall performance of the system. It also helps in reducing the idle time of the computing nodes hence improving their utilization [16].

The Compare and Balance algorithm uses the concept of compare and balance to reach an equilibrium condition and manage unbalanced system's load. On the basis of probability (no. of virtual machine running on the current host and whole cloud system), current host randomly select a host and compare their load. If load of current host is more than the selected host, it transfers extra load to that node. Each host of the system performs the same procedure [2]. This algorithm assures that the migration of VMs is always from high cost physical hosts to low-cost host but assumes that each physical host has enough memory which is a weak assumption [7].

The Carton is a technique that is a combination of Load Balancing (LB) and Distributed Rate Limiting (DRL). LB (Load Balancing) is used to equally distribute the jobs to different servers so that the associated costs can be minimized. DRL ensure the equal distribution of resources. Work load is dynamically assigned to improve the performance and spread the load equally to all the servers. With very low computation and communication overhead, this algorithm is simple and easy to implement [7].

The Cloud Partitioning Concept is a dynamic load balancing technique for cloud based on partitioning concept with a switch mechanism to choose different strategies for different situations. Cloud Partitioning is for Public cloud which has numerous nodes with distributed computing resources at different geographic locations. Here cloud divides into several cloud partitions. The cloud has main controller that chooses the suitable partitions for arriving jobs and there is balancer for cloud partition which chooses the correct load balancing strategy. If the load status of a cloud partition is idle and normal, this partitioning accomplished job locally and when load status is not normal, another cloud partition is searched [9].

VI. CONCLUSION

This paper discussed cloud computing. In cloud computing, there are infinite computing capabilities with attractive pay-per-use scheme. One of the major issues of cloud computing is system load balancing, because overloading of a particular node makes it slow down resulting poor system efficiency. So there is always a requirement of efficient load balancing algorithms for improving the utilization of computing resource. In this paper, we have surveyed various load balancing techniques for cloud computing. The main purpose of load balancing is to satisfy the customer requirement by distributing load dynamically among the nodes and to make maximum resource utilization by reassigning the total load to individual node. This ensures that every resource is distributed efficiently and evenly. So the performance of the system is increased.

REFERENCES

1. Bala K. et. al., 'A Review Of The Load Balancing Techniques At Cloud Server', International Journal of Advances in Computer Science and Communication Engineering (IJACSCE), Vol. 2, Issue I, 2014, pp 6-11.
2. Desai T. and Prajapati J., 'A Survey of Various Load Balancing Techniques and Challenges in Cloud Computing', International Journal Of Scientific & Technology Research, Vol. 2 , Issue 11, 2013, pp 158-161.
3. Katyal M. and Mishra A., 'A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment', International Journal of Distributed and Cloud Computing, Vol. 1, Issue 2, 2013, pp 5-14.
4. Sahu Y. and Pateriya R., 'Cloud Computing Overview with Load Balancing Techniques', International Journal of Computer Applications , Vol. 65– No.24 , 2013, pp 40-44.
5. Gabi D. et. al., 'Systematic Review on Existing Load Balancing Techniques in Cloud Computing', International Journal of Computer Applications, Vol. 125 – No.9, 2015, pp 16-24.
6. Palta R. and Jeet R., 'Load Balancing in the Cloud Computing Using virtual Machine Migration: A Review', International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 3, Issue 5, 2014, pp 437-441.
7. Ramesh B.,' Load balancing in Cloud Computing – An Analysis', International Conference on Security and Authentication - SAPIENCE141, 2014, pp 125-131.
8. Gopinath G. and Vasudevan S., 'An in-depth analysis and study of Load balancing techniques in the cloud computing environment', 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), 2014, pp 427-432.
9. More S. and Mohanpurkar A., 'Load Balancing Strategy Based On Cloud Partitioning Concept',

Multidisciplinary Journal of Research in Engineering and Technology, Vol. 2, Issue 2, 2015, pp 424-431.

10. Kaur P. and Kaur P., 'Efficient and Enhanced Load Balancing Algorithms in Cloud Computing', International Journal of Grid Distribution Computing, Vol. 8, No.2 , 2015 pp 9-14.
11. Panwar R. and Mallick B., 'A Comparative Study of Load Balancing Algorithms in Cloud Computing', International Journal of Computer Applications, Vol. 117 – No. 24, 2015, pp 33-37.
12. Sadhu R. and Vania J., 'Survey on Various Load Balancing Techniques in Cloud Computing Environment', International Journal of Science and Research (IJSR), Vol. 4, Issue 10, 2015, pp 1881-1884.
13. Vouk M., 'Cloud Computing – Issues, Research and Implementations', Journal of Computing and Information Technology – CIT, 2008, pp 235–246.
14. Megharaj G. and Mohan K.G., 'Two Level Hierarchical Model of Load Balancing in Cloud', International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 10, 2013, pp 307-311.
15. Singh A. et. al., 'Comparative Analysis Of Load Balancing Algorithms In Cloud Computing', International Journal of Advanced Technology & Engineering Research (IJATER), Vol. 4, Issue 2, 2014, pp 18-21.
16. Gupta R., 'Review on Existing Load Balancing Techniques of Cloud Computing', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 2, 2014, pp 168-171.
17. Pasha N. et. al., 'Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 5, 2014, pp 34-39.
18. Yadav D. and Keswani B., 'Porting Intranet Over Cloud for Educational Service Amplification (Special Reference to Higher Educational Institutions)', SYLWAN, Vol. 161, Issue 8, 2017.
19. Sharma R. and Keswani B., 'Study & analysis of cloud based ERP services', International Journal of Mechatronics, Electrical and Computer Technology, Vol. 3, Issue 9, 2013, pp. 375-396.
20. Yadav D. and Keswani B., 'A Study of Intranet over Cloud', International Journal of New Innovations in Engineering and Technology, Vol. 7, Issue 2, 2017, pp. 1-6.
21. Ikhlaq S. and Keswani B., Computation of Big Data in Hadoop and Cloud Environment', IOSR Journal of Engineering (IOSRJEN), Vol. 6, Issue 1, 2016, pp. 31-39.